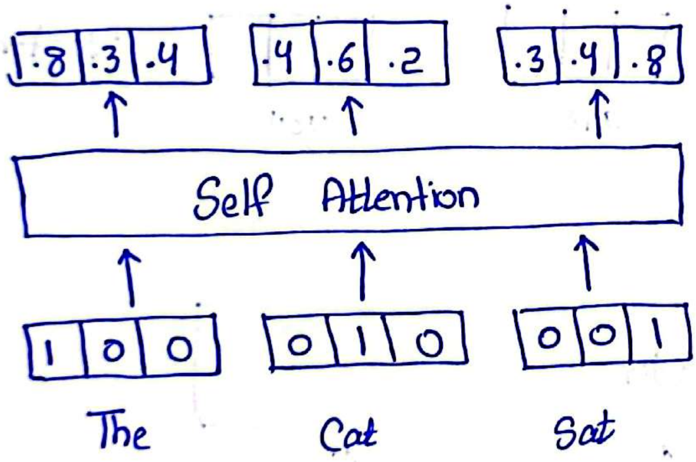
→ Self-Attention

Self-attention, also known as scaled dot-product attention, is a crucial mechanism in the transformer architecture that allows the model to weigh the importance of different tokens in the input sequence relative to each other

> Idea



> Steps

1) Inputs

Q: Queries

K: Keys

V: Values

Model → Q, K, V

Ø

- Query Vector (Q)

Query vector represent the token for which we are calculating the attention. They help determine the importance of other tokens in the context of current token

Importance

⊢ Focus Determination

Queries help the model to decide which parts of the sequence to focus on for each specific token. By calculating the dot product between a query vector (Q) and all key vectors (K), the model assesses how much attention to give to each token relative to the current token

⊢ Contextual Understanding

Queries contribute to understanding the relationship between the current token and the rest of the sequence, which is essential for capturing dependencies and context.

- **Key Vectors (K):**

Key vectors represent all the token in the sequence and are used to compare with the query vectors to calculate attention scores

Importance

⊢ **Relevance Measurement**

Keys are compared with queries to measure the relevance or compatibility of each token with the current token. This comparison helps in determining how much attention each token should have

⊢ **Information Retrieval**

Keys plays a critical role in retrieving the most relevant information from the sequence by providing a basis for the attention mechanism to compute similarity scores.

- **Value Vectors (V):**

Value vectors holds the actual information that will be aggregated to form the output of the attention mechanism

Importance

⊢ Information Aggregation

Values contain the data that will be weighted by the attention scores. The weighted sum of values forms the output of the self-attention mechanism, which is then passed on to the next layers in the network.

⊢ Context Preservation

By weighting the values according to the attention scores, the model preserves and aggregates relevant context from the entire sequence, which is crucial for tasks like translation, summarization and more.

- **Example**

Input sequence = ["The", "Cat", "Sat"]

Embedding size = 4 dimensions.

Q, K, V = 4 dimensions

① <u>Token Embeddings</u>

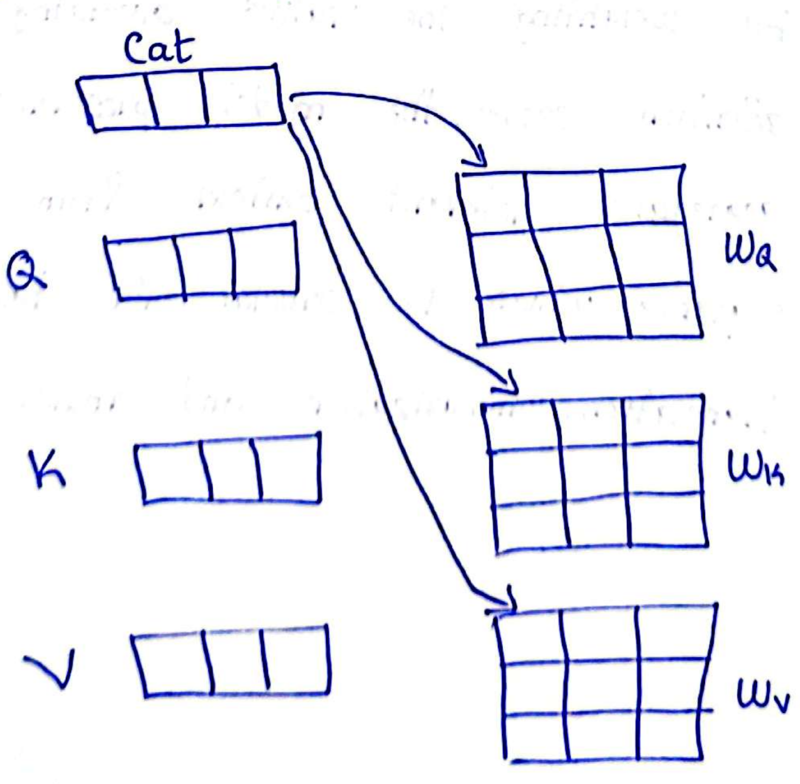
$E_{The} = [1, 0, 1, 0]$

$E_{cat} = [0, 1, 0, 1]$

$E_{sat} = [1, 1, 1, 1]$

② <u>Linear Transformation</u>

We create Q, K, V by multiplying the embeddings by learned weight matrices : W_Q, W_K, W_V



$$W_Q = W_K = W_V = I \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$Q_{The} = E_{The} \cdot W_Q$$

$$Q_{The} = [1\ 0\ 1\ 0]\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = [1\ 0\ 1\ 0]$$

$$K_{The} = [1\ 0\ 1\ 0] \quad \because E_{The} \cdot W_K$$

$$V_{The} = [1\ 0\ 1\ 0] \quad \because E_{The} \cdot W_V$$

1- $Q_{The} = K_{The} = V_{The} = [1\ 0\ 1\ 0]$

2- $Q_{cat} = K_{cat} = V_{cat} = [0\ 1\ 0\ 1]$

3- $Q_{sat} = K_{sat} = V_{sat} = [1\ 1\ 1\ 1]$

3 Compute Attention Scores

For token = the

$$\text{Score}(Q_{The}, K_{The}) = [1\ 0\ 1\ 0][1\ 0\ 1\ 0]^T = 2$$

$$\text{Score}(Q_{The}, K_{cat}) = [1\ 0\ 1\ 0][0\ 1\ 0\ 1]^T = 0$$

$$\text{Score}(Q_{The}, K_{sat}) = [1\ 0\ 1\ 0][1\ 1\ 1\ 1]^T = 2$$

For token = cat

$$\text{Score}(Q_{cat}, K_{The}) = [0\ 1\ 0\ 1][1\ 0\ 1\ 0]^T = 0$$

$$\text{Score}(Q_{cat}, K_{cat}) = [0\ 1\ 0\ 1][0\ 1\ 0\ 1]^T = 2$$

$$\text{Score}(Q_{cat}, K_{sat}) = [0\ 1\ 0\ 1][1\ 1\ 1\ 1]^T = 2$$

For token = sat

$$\text{Score}(Q_{sat}, K_{The}) = [1\ 1\ 1\ 1][1\ 0\ 1\ 0]^T = 2$$

$$\text{Score}(Q_{sat}, K_{cat}) = [1\ 1\ 1\ 1][0\ 1\ 0\ 1]^T = 2$$

$$\text{Score}(Q_{sat}, K_{sat}) = [1\ 1\ 1\ 1][1\ 1\ 1\ 1]^T = 4$$

4 Scaling

We take the scores and scale down by dividing the scores by the square root of dimensions of key vector

$$d_k = \sqrt{4}$$

$$= 2$$

Scaling in attention mechanism is crucial to prevent the dot product from growing too large

Problems:

1- Gradient Exploding

2- Softmax Saturation

Example without scaling

$$Q = [2 \quad 3 \quad 4 \quad 1] \qquad K_1 = [1 \quad 0 \quad 1 \quad 0] \qquad K_2 = [0 \quad 1 \quad 0 \quad 1]$$

$$Q.K_1^T = 2\times1 + 3\times0 + 4\times1 + 1\times0 = 6$$

$$Q.K_2^T = 2\times0 + 3\times1 + 4\times0 + 1\times1 = 4$$

Score : $[6, 4]$

$$\text{Softmax}([6,4]) = \left[\frac{e^6}{e^6+e^4}, \frac{e^4}{e^6+e^4}\right]$$

$$= \left[\frac{e^6}{e^6(1+e^{-2})}, \frac{e^4}{e^4(e^2+1)}\right]$$

$$= \left[\frac{1}{1+e^{-2}}, \frac{1}{e^2+1}\right] \approx [0.88, 0.12]$$

difference is huge

Most of the attention weights are assigned to the first vee Key vector

Example with Scaling

Score : $[6, 4]$ $\Rightarrow$ Scale $\Rightarrow$ $[6/2, 4/2] = [3, 2]$

$$\text{Softmax}([3,2]) = \left[\frac{e^3}{e^3+e^2}, \frac{e^2}{e^3+e^2}\right] = \left[\frac{e^3}{e^3(1+e^{-1})}, \frac{e^2}{e^2(e^1+1)}\right]$$

$$= \left[\frac{1}{1+e^{-1}}, \frac{1}{e^1+1}\right] \approx [0.73, 0.27]$$

Distance is less difference

④ Scaling Cont.

for token: the

$$\text{Scaled-score}(Q_{The}, K_{The}) = 2/2 = 1$$

$$\text{Scaled-score}(Q_{The}, K_{cat}) = 0/2 = 0$$

$$\text{Scaled-score}(Q_{The}, K_{sat}) = 2/2 = 1$$

for token = cat

$$\text{Scaled-score}(Q_{cat}, K_{The}) = 0/2 = 0$$

$$\text{Scaled-score}(Q_{cat}, K_{cat}) = 2/2 = 1$$

$$\text{Scaled-score}(Q_{cat}, K_{sat}) = 2/2 = 1$$

for token = sat

$$\text{Scaled-score}(Q_{sat}, K_{The}) = 2/2 = 1$$

$$\text{Scaled-score}(Q_{sat}, K_{cat}) = 2/2 = 1$$

$$\text{Scaled-score}(Q_{sat}, K_{sat}) = 4/2 = 2$$

⑤ Apply Softmax

$$\text{Attention Weight}_{The} = \text{Softmax}([1,0,1]) = [0.42, 0.15, 0.42]$$

$$\text{Attention Weight}_{cat} = \text{Softmax}([0,1,1]) = [0.15, 0.42, 0.42]$$

$$\text{Attention Weight}_{sat} = \text{Softmax}([1,1,2]) = [0.21, 0.21, 0.58]$$

⑥ Weighted Sum of Values

Multiply attention weights to corresponding value vector

$$Output_{(The)} = 0.42 \times V_{The} + 0.15 \times V_{cat} + 0.42 \times V_{sat}$$

$$= 0.42[1\ 0\ 1\ 0] + 0.15[0\ 1\ 0\ 1] + 0.42[1\ 1\ 1\ 1]$$

$$= [0.42\ \ 0\ \ 0.42\ \ 0] + [0\ \ 0.15\ \ 0\ \ 0.15] + [0.42\ \ 0.42\ \ 0.42\ \ 0.42]$$

$$= [0.84\ \ 0.57\ \ 0.84\ \ 0.57]$$

$$[1\ \ 0\ \ 1\ \ 0] \Rightarrow \text{Self-Attention} \Rightarrow [0.84\ \ 0.57\ \ 0.84\ \ 0.57]$$

$$Output_{(cat)} = 0.15 \times V_{The} + 0.42 \times V_{cat} + 0.42 \times V_{sat}$$

$$= 0.15[1\ 0\ 1\ 0] + 0.42[0\ 1\ 0\ 1] + 0.42[1\ 1\ 1\ 1]$$

$$= [0.15\ \ 0\ \ 0.15\ \ 0] + [0\ \ 0.42\ \ 0\ \ 0.42] + [0.42\ \ 0.42\ \ 0.42\ \ 0.42]$$

$$= [0.57\ \ 0.84\ \ 0.57\ \ 0.84]$$

$$Output_{(sat)} = 0.21[1\ 0\ 1\ 0] + 0.21[0\ 1\ 0\ 1] + 0.58[1\ 1\ 1\ 1]$$

$$= [0.21\ \ 0\ \ 0.21\ \ 0] + [0\ \ 0.21\ \ 0\ \ 0.21] + [0.58\ \ 0.58\ \ 0.58\ \ 0.58]$$

$$= [0.79\ \ 0.79\ \ 0.79\ \ 0.79]$$